Towards Targeted Sales

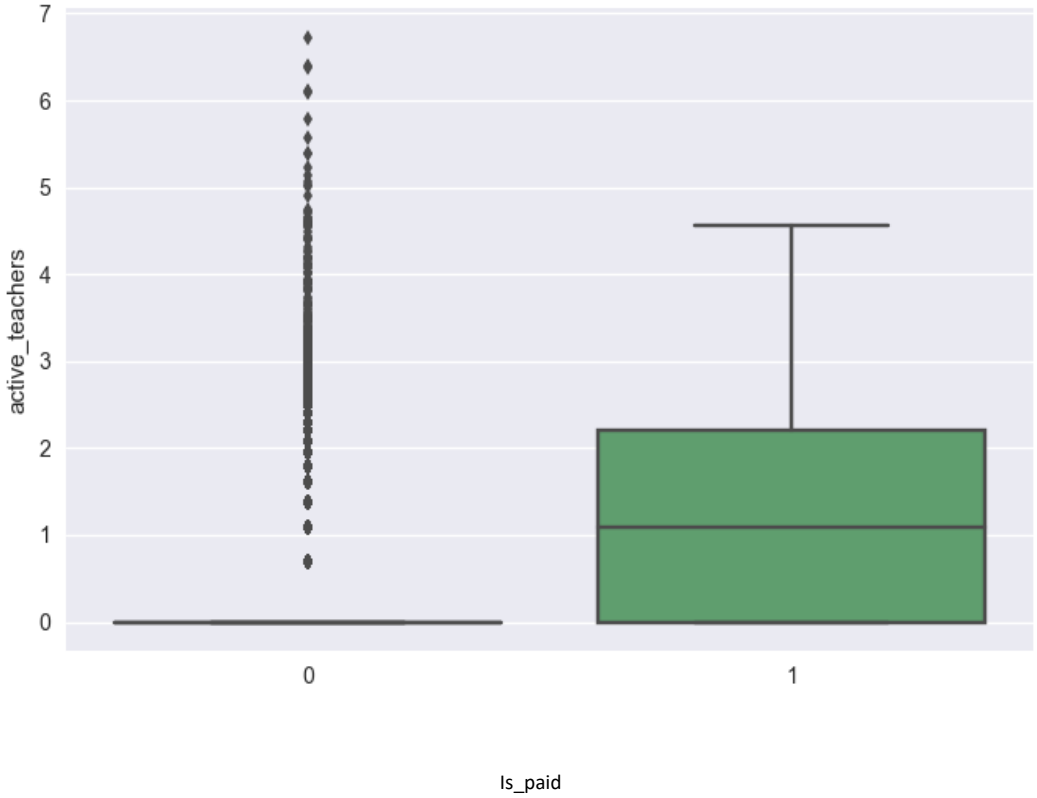
Currently, sales leads are pursued primarily on number of activated teachers in the school. While this is an elegantly simple way to target potential leads, a rich set of available data makes possible the ability to determine the probability of sign-up on an individual basis & facilitate more precise targeting of likely sales.

Assessing variables

Below table shows performance of the top 10 variables in predicting target variable "is_paid".

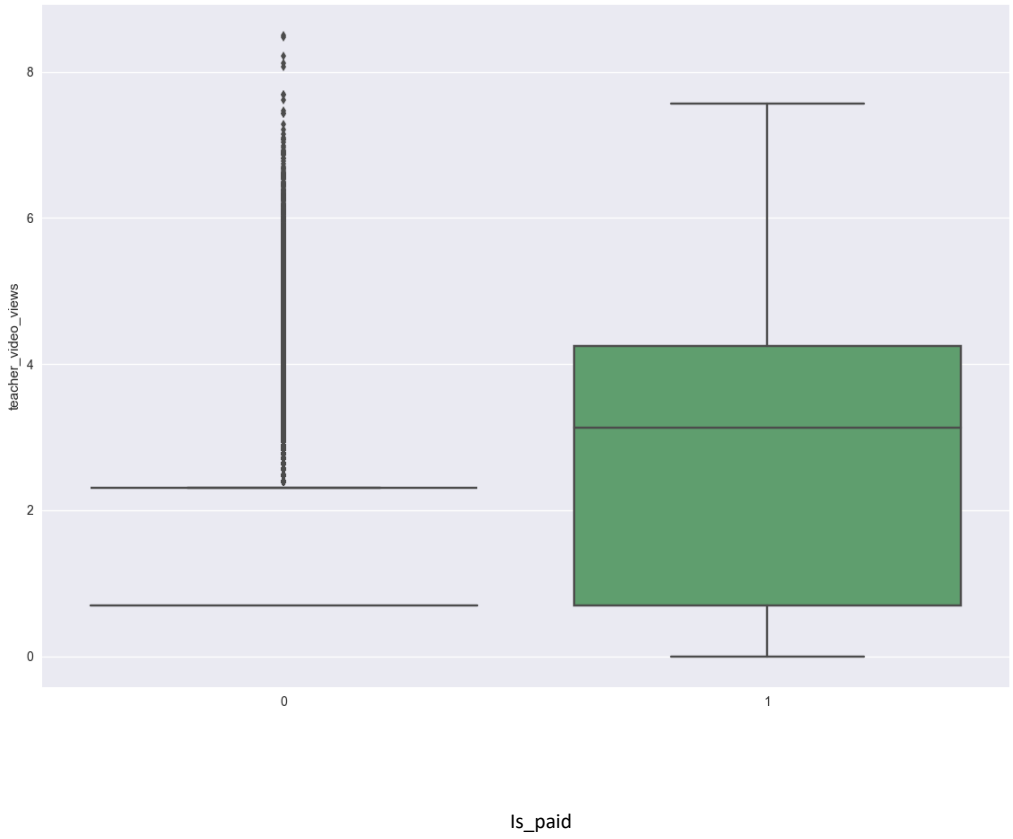
Variable	**F-Score**
teacher_video_views	4273.637
w4_active_teachers	3605.24
active_teachers	3581.712
w5_active_teachers	3188.656
w4_teacher_video_views	3101.856
w3_active_teachers	3053.746
w2_active_teachers	2892.311
w1_active_teachers	2742.88
w5_teacher_video_views	2604.503
w2_teacher_video_views	2562.684

Higher F-scores are indicative of stronger relationships with the target variable. Based on these scores, it is clear that number of active teachers is highly indicative of whether or not the school has a paid subscription—so as a simple way to target potential sales leads, is a useful indicator.



Above plot displays distribution of values “active_teachers” (log-transformed, to better display distribution of values) versus the “is_paid” variable. The plot shows a distinctly higher distribution of number of active teachers where there is a paid subscription, compared to where there is not. The area of the green box for “is_paid” = 1 shows that the middle 50% of values (those closest to median) are concentrated at a much higher level than the analogous range where “is_paid” = 0. The data tells us that where there are more active teachers, the chances of paid subscription are indeed higher.

Based on the scoring data, a variable that receives a higher score than “active_teachers” is “teacher_video_views”. Let us examine the variable more closely.



Above plot displays distribution of “teacher_video_views” (again, log-transformed) against the “is_paid” indicator. From above plot, it is clear that paid subscriptions have a distinctly higher median & concentration of values than non-paid subscriptions. While the overlap of values does not seem as distinct as with “active_teachers”, activity of “teacher_video_views” does seem distinct enough so as to differentiate activity between paid & non-paid users—as such, we can presume that higher video views indicates higher probability of a paid subscription.

Other questions

- Are there other potentially actionable insights to be gleaned from this dataset?

Above analysis examines the relationships between the variables and potential sales leads. It is apparent from this analysis that higher engagement with the platform (as measured by number of users within a school, or video views) tends to lead to a sale. The next area of exploration should be to determine the drivers of user engagement, which may lead eventually to a sale. If the reasons behind, video views can be determined, for example, then users can be targeted more directly to encourage video views—which may in turn lead to higher sales.

● How would you communicate these results to the sales team?

The ideal way to communicate these results to the sales team would be some kind of dashboard-based tool such as Tableau or Looker, which visualizes the major indicators (such as those described above) of promising sales leads, and allows their team to "drill down" into the indicators to explore users on a more individual basis. Additionally, I'd make the list of most likely converters (as determined by predictive score) available through the dashboard as well; envisaging that the team is structured according to regional focus, each team could potentially have a dashboard screen configured by regional area of interest, but which is based on one large data set, to avoid repeating work at BI team level.

● How would you operationalize this analysis so that it is repeatable and accurate?

The model would be maintained in some sort of re-runnable script (likely Python-based), which would rely on a periodic data refresh from another system (theoretically SQL-based). The Python script would be re-run on a periodic basis (likely monthly) to incorporate latest data and measure accuracy of information against previous periods. Major dips in accuracy would likely trigger a re-analysis of the data and tune-up of the model. Another option to operationalize a model may be a data science platform like Dataiku, which is designed for team collaboration and deployment in the development of data science models that is not necessarily based on scripting languages less intuitive to non-data scientists.